Information Retrieval from Multiple Sources Using a Simple Web Interface

Phyllis A. Koton, Ph.D., Rick Lawson, MSLIS, John P. LaFond, William S. Reece
HealthGate Data Corp., Malden, MA

We have developed a system for Internet information retrieval that uses a uniform interface to access information from multiple sources, including bibliographic databases, full-text journals, reference books and consumer health reports. The intent in designing the system was to provide access to a vast amount of medical information to the non-professional searcher, without requiring knowledge of specific vocabularies, command languages or database contents.

To use the system, the user enters a query in natural language. The question is analyzed by matching sentence fragments against a set of patterns representing concepts, relationships between concepts, and types of articles. Synonyms and lexical variants are converted to controlled-vocabulary terms for each data source, when such a vocabulary exists for the source. Domain-specific constraints are used to select a single interpretation from multiple matching patterns, if necessary. The user's query is reformulated into a search strategy for the appropriate information source and the results are presented to the user.

For example, if the user enters "can taxol be used to treat breast cancer" and one of the available information sources uses the National Library of Medicine's Medical Subject Headings (MeSH) indexing terms, the system will formulate the query "paclitaxel and breast-neoplasms-DT". The sentence fragment "be used to treat" identifies the *is-therapy* concept. This concept requires the presence of a disorder and a therapy. In this query, the fragment "breast cancer" fills the disorder role and is translated to the MeSH term "breast neoplasms". The fragment "taxol" fills the therapy role, and is translated to the MeSH term "paclitaxel". Because taxol is a drug, the therapy concept is translated as the subheading DT. The system also checks to make sure that the MeSH term "breast neoplasms" can use the subheading DT. Finally the system adds codes that indicate in which field of the database to look for this query string. For example, if the database had a MeSH Major Heading field, the code for this field would be added.

A similar process is used to translate the query for each data source. For data sources that do not have specific vocabularies, the query is processed to eliminate excess words (such as "find articles" or "tell me about") and stopwords. The search is then processed and the results are returned.

The advantages of an interface such as this one for the non-professional users are: (1) no knowledge of the controlled vocabularies for each database is required, (2) no knowledge of database fields is required, (3) the user only has to learn a single interface for searching all files. The system has been running on the World Wide Web since January 1996 and reaction from users has been overwhelmingly positive.

We are currently working on improving the search strategies formulated by the system in two main ways. We plan to use the UMLS Semantic Network to identify semantic types and disambiguate relationships between concepts, as described in Cimino et al[1]. We are currently implementing question categories (Burke et al[2]) to generate search strategies for more general questions such as "What are ..." and "How often should I ..." whose posers are often overwhelmed by the results of a keyword search on bibliographic databases. Question categories will also allow us to better match the user's queries with available information sources.

References

1. Cimino JJ, Aguirre A, Johnson SB, Peng P. Generic queries for meeting clinical information needs. Bull Med Libr Assoc, 1993:81(2):195-206.
2. Burke R, Hammond K, Kozlovsky J. Knowledge-based Information Retrieval from Semi-Structured Text. AAAI-95 Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, November 10-12. Cambridge, MA. AAAI, 1995.

0195—4210/96/$5.00 © 1996 AMIA, Inc.